Chapter 9
Materials Data Infrastructure and Materials Informatics

Joanne Hill, Arun Mannodi-Kanakkithodi, Ramamurthy Ramprasad, and Bryce Meredig

9.1 Materials Data Infrastructure

The materials science and engineering (MS&E) community identified its collective need for data infrastructure as early as the 1980s [1]. While interest around this topic has grown markedly in recent years, there is still much work to do in order to meet the field's needs and unlock the potential benefits of advanced data infrastructure. Data infrastructure plays an important role for fundamental materials researchers as well as materials producers and materials-enabled product companies (MEPCs or manufacturers whose products rely on advanced materials) as data generated through laboratory investigations, the manufacturing environment, and customer specifications need to be stored in an easily searchable, updateable, and accessible infrastructure [3].

Currently, there is a drive to develop materials databases built around highly structured processing and property relationships, as these will enable materials informatics techniques and could assist with materials discovery, development, and deployment [4–9]. The Materials Genome Initiative (MGI) [6] and the 2013 open access memo from the White House Office of Science and Technology Policy [10] are two examples of US-based efforts that are encouraging the community to work toward solutions to meet the current infrastructure need. In addition to US efforts, there are other international projects, such as the European Commission Joint Research Centre's MatDB [11] and the European Union's NoMaD repository

J. Hill • B. Meredig (✉)
Citrine Informatics, Redwood City, CA 94062, USA
e-mail: jo@citrine.io; bryce@citrine.io

A. Mannodi-Kanakkithodi • R. Ramprasad
Department of Materials Science and Engineering, University of Connecticut, Storrs, CT 06269, USA
e-mail: mannodiarun@gmail.com; rampi.ramprasad@uconn.edu

D. Shin, J. Saal (eds.), *Computational Materials System Design*,
https://doi.org/10.1007/978-3-319-68280-8_9

[12] that have been developed to address infrastructure requirements. The National Institute of Materials Science in Japan is also looking at ways to extend their existing databases to address this issue [5]. Table 9.1 is an extensive, yet inevitably incomplete, list of the currently available materials data resources from around the world [13].

As can be seen in Table 9.1, current infrastructure efforts in this field follow one of two approaches. These are either (1) creating general databases that focus on storing as much data as possible without imposing rigid structure or file format restrictions such as National Institute of Standards and Technology's (NIST) DSpace repository [14] and the Materials Data Facility [15] or (2) creating databases that are specific to a certain subdiscipline of materials and designed to store only domain-specific information. The Inorganic Crystal Structure Database (ICSD) [16] is one example of this type of database [2, 17].

9.1.1 Data Infrastructure Considerations

Materials data infrastructure comprises complex software involving many design choices. A general architecture can be thought of in terms of three key features: (1) data import, (2) data storage, and (3) data access. We explore these features in detail below.

9.1.1.1 Data Import

Materials data are currently stored in a large number of different ways, dependent on factors such as the area of research, data source, acquisition technique, and type of post-processing. On one hand, ideally a materials infrastructure should be able to support as many different data formats as possible and import documents regardless of type. However, data are vastly more useful when they are structured in a way that allows systematic search and analysis of the content, which is inherently more difficult with greater diversity of data. This means that along with importability, an ideal infrastructure should be able to structure uploaded data in a standardized way to as great an extent as possible [2].

The ability to incorporate new data in real time is also becoming increasingly important as data-intensive analysis approaches, such as materials informatics, benefit from a continual stream of new knowledge [18]. In order for this to be possible, there needs to be support for programmatic upload and download of data, such as through an application program interface (API).

Table 9.1 A list of some notable materials data resources from [13]

Name	URL	Category	Free/non-free
3D Materials Atlas	cosmicweb.mse.iastate.edu/wiki/display/home/Materials+Atlas+Home	3D characterization	Free
AFLOWLIB	aflowlib.org	Computational	Free
AIST research information databases	www.aist.go.jp/aist_e/list/database/riodb	General materials data	Free
American Mineralogist Crystal Structure Database	rruff.geo.arizona.edu/AMS/amcsd.phP	Minerals	Free
ASM Alloy Center Database	mio.asminternational.org/ac	Alloys	Non-free
ASM Phase Diagrams	wwwl .asminternational.org/AsmEnterprise/APD	Thermodynamics	Non-free
CALPHAD databases (e.g., Thermocalc SGTE)	www.thermocalc.com/products-services/databases/thermodynamic	Thermodynamics	Non-free
Cambridge Crystallographic Data Centre (CCDC)	www.ccdc.cam.ac.uk/pages/Home.aspx	Crystallography	Non-free
CatApp	suncat.stanford.edu/catapp	Catalysts	Free
ChemSpider	wvrtv.chemspider.com	Chemical data	Free
CINDAS High-Performance Alloys Database	cindasdata.com/products/hpad	Alloys	Non-free
Citrination	citrination.com	General materials data	Free
Computational Materials Repository	cmr.fysik.dtu.dk	Computational	Free
CRC Handbook	www.hbcpnetbase.com	General materials data	Non-free
CrystMet	cds.dl.ac.uk/cgi-bin/news/disp?crystmet	Crystallography	Non-free
Crystallography Open Database (CoD)	http://www.crystallography.net	Crystallography	Free
DOE Hydrogen Storage Materials Database	www.hydrogenmaterialssearch.govtools.us	Hydrogen storage	Free
Granta CES Selector	www.grantadesign.com/products/ces	General materials data	Non-free
Handbook of Optical Constants of Solids, Palik	N/A	Hard-copy sources	Non-free

(continued)

Table 9.1 (continued)

Name	URL	Category	Free/non-free
Harvard Clean Energy Project	cepdb.molecularspace.org	Computational	Free
Inorganic Crystal Structure Database (ICSD)	cds.dl.ac.uk/cds/datasets/crys/icsd/llicsd.html	Crystallography	Non-free
International Glass Database System (INTERGLAD)	wvrtv.newglass.jp/interglad_n/gaiyo/info_e.html	Glass	Non-free
Knovel	app.knovel.com/web/browse.v	General materials data	Non-free
Matbase	www.matbase.com	General materials data	Free
MatDat	www.matdat.com	General materials data	Non-free
Materials Project	wvrtv.materialsproject.org	Computational	Free
MatNavi (NIMS)	mits.nims.go.jp/index_en.html	General materials data	Free
MatWeb	www.matweb.com	General materials data	Free
Mindat	www.mindat.org	Minerals	Free
NanoHUB	nanohub.org	Nanomaterials	Free
Nanomaterial Registry	www.nanomaterialregistry.org	Nanomaterials	Free
NIST Materials Data Repository (DSpace)	materialsdata.nist.gov/dspace/xmlui	General materials data	Free
NIST Interatomic Potentials Repository	vrtvw.ctcms.nist.gov/potentials	Computational	Free
NIST Standard Reference Data	www.nist.gov/srd/dblistpcdatabases.cfm	General materials data	Non-free
NIST Standard Reference Data	www.nist.gov/srd/onlinelist.cfm	General materials data	Free
NoMaD	nomad-repos itory.eu/cms	Computational	Free
Open Knowledge Database of Interatomic Models (OpenKIM)	openkim.org	Computational	Free
Open Quantum Materials Database	oqmd.org	Computational	Free

Pauling File	paulingfile.com	General materials data	Non-free
Pearson's Handbook: Crystallographic Data	N/A	Hard-copy sources	Non-free
Powder Diffraction File (PDF)	www.icdd.com/products/index.htm	Crystallography	Non-free
PubChem	pubchem.ncbi.nlm.nih.gov	Chemical data	Free
Reaxys	www.elsevier.com/solutions/reaxys	Chemical data	Non-free
SciFinder/ChemAbstracts	scifinder.cas.org	Chemical data	Non-free
SciGlass	www.sciglass.info	Glass	Non-free
SpringerMaterials	materials.springer.com	General materials data	Non-free
Metallurgical Thermochemistry, Kubaschewski	N/A	Hard-copy sources	Non-free
TEDesignLab	www.tedesignlab.org	Thermoelectrics	Free
Total Materia	www.totalmateria.com	General materials data	Non-free
UCSB-MRL thermoelectric database	www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp	Thermoelectrics	Free

9.1.1.2 Data Storage

There are many different technologies available for data storage, each with its own advantages and disadvantages. As there is no ideal solution that suits all needs, there are three main factors that need to be considered when choosing how data should be stored. First is the structure of the data. This is important as the storage mechanism used can impact the way in which the data can be stored, imported, and retrieved. Second is the access pattern, including how often the data will be accessed, where they will be accessed from, and what security the data require. Finally, the storage mechanism must meet operational requirements for query performance, security, data availability, and scalability. Many architectures make use of multiple technologies. This allows for a custom solution that best suits the application. Relational databases, non-relational databases, and object stores are three examples of commonly used storage technologies [2].

Relational databases are well suited to data that fits into a relational model, i.e., data that can be stored in tables with columns and rows. Access to data stored in this way is very fast, which makes it well suited to storing information that needs to be returned quickly. User account information, session information, and user authentication information are some examples of data that can be application critical, necessitating short query times [19, 20].

Non-relational databases are used to store data that does not fit well into a traditional relational database. It differs from a relational database in that it allows any record to be accessed as long as the record key is known [21]. This method also allows rapid access to data, and materials scientists are starting to see the benefits of non-relational data storage [22].

Object stores, which manage data as objects, are scalable and resilient but have a much slower response time and are thus not well suited to storing data that needs to be accessed often or quickly. However, even though access is slow, object stores can be useful for saving ground truth data that is very important but need not be accessed often [23].

9.1.1.3 Data Access

Appropriate data access methods need to be selected based on the use case for the infrastructure and data. In some instances, a user may simply need the original documents returned, while in others they may need structured data in order to perform complex queries or access specific data points from a larger dataset. Access methods will also differ depending on whether the users require programmatic access to the data or simply a graphical interface that allows them to search for data of interest and download the required information. APIs are commonly used for programmatic access to data, and many user interfaces are simply tools that allow graphical access to data returned by the API [2].

9.1.2 Data Standards

Along with the need for infrastructure and databases in which to store materials data, there is also a critical need for widely accepted data standards. There are currently a number of different schemas that have been developed to store materials data, but these are generally very narrowly focused on a single subdiscipline within the MS&E field. One example of such a schema is the Crystallographic Information File (CIF) [24]. This format has become the gold standard for storing crystallographic data, and it performs this function very well. However, it is very rigidly structured and thus not suitable for storing any data that are more generalized. More general schemas are starting to emerge for the materials community, but as yet there are not any that have been widely adopted [25–28]. This may be due, at least in part, to the inherent difficulty involved in creating a suitable schema. The schema needs to be general enough to represent the wide variety of data that is generated in MS&E and flexible enough to store data that may not have originally been considered by its developers while not being *so* general that users realize no tangible benefits from its adoption.

Different organizations have addressed and dealt with these requirements in a number of different ways. Next, we take a look at four distinctive infrastructure implementations from different groups.

9.1.3 Data Infrastructure: Citrination Platform

Citrine Informatics has taken an approach to data infrastructure that attempts to balance the need for a general database that can incorporate cross-disciplinary data from all areas of MS&E with the need to structure data in a way that makes it easily accessible, searchable, and machine readable [2]. The Citrination platform [29] is a single centralized location for data from various fields in the materials space and can be used to store, access, and analyze structured and unstructured data within a cloud-based infrastructure. Citrination stores data with the goal of making sure they are both human searchable and machine readable for the purpose of algorithmic data mining [2]. By consolidating and structuring data within a single infrastructure, Citrine is able to readily use these data as training examples for machine learning and other modeling techniques [18]. Further, Citrine expects that enabling cross-pollination across disciplines and easy access to data will enable advancements in materials that would not otherwise be possible [13].

9.1.4 Data Standards: The Physical Information File

With predictably structured data being the most important prerequisite for the successful implementation of materials informatics, Citrine has developed a hierarchical data structure for storing materials data, called the Physical Information File (PIF). This is an open-source, machine-readable structured format that can accommodate complex materials data [30, 31].

The PIF was designed to represent very diverse data related to materials and physical systems, ranging from the atomic to the macro scale, describing these objects' processing history and properties. The PIF also allows all of the information from a system to be contained in a single file. For example, a file can contain information about a wing for a plane, the parts that make up the wing and the properties of the materials from which those parts are made. An example of this can be seen in Fig. 9.1 [30].

The primary design goal for the PIF was to make it suitable for storing vastly different types of materials data without making it too difficult to adopt or understand [30]. The language to describe metals, alloys, and polymers differs widely, but there is no reason that we should not be able to compare these classes of materials on properties that are common to all, such as yield strength or toughness; using a single-file format to store information on all of these materials greatly simplifies this task [32]. The PIF cannot replace all data files for physical systems, but it does provide a suitable way to store a broad variety of system-material-processing-property information [30].

In order to enable the breadth and flexibility required in the PIF, we must also allow for some ambiguity. For example, "heat of formation" and "formation enthalpy" may occur separately in a single PIF record, while it could be argued that these are the same concepts. However, as the data is machine readable, algorithms and heuristics can be used to organize the data in a meaningful way for the required application. For example, we could deploy software to make the determination that given conceptual overlap within PIFs and other documents, numerical similarity, unit similarity, heat of formation, and formation enthalpy are indeed synonymous. Citrine Informatics has developed open-source tools for building files that use the PIF schema and working with PIFs. This assists in lowering the barriers to use for this data format [30].

9.1.5 Citrination Platform Architecture

The Citrination platform was developed to enable the entire materials community to use a single cohesive data infrastructure for storing their research outputs. We illustrate the basic architecture of Citrination in Fig. 9.2 [2]. The key design choices underpinning the architecture relate to data import, storage, and access.

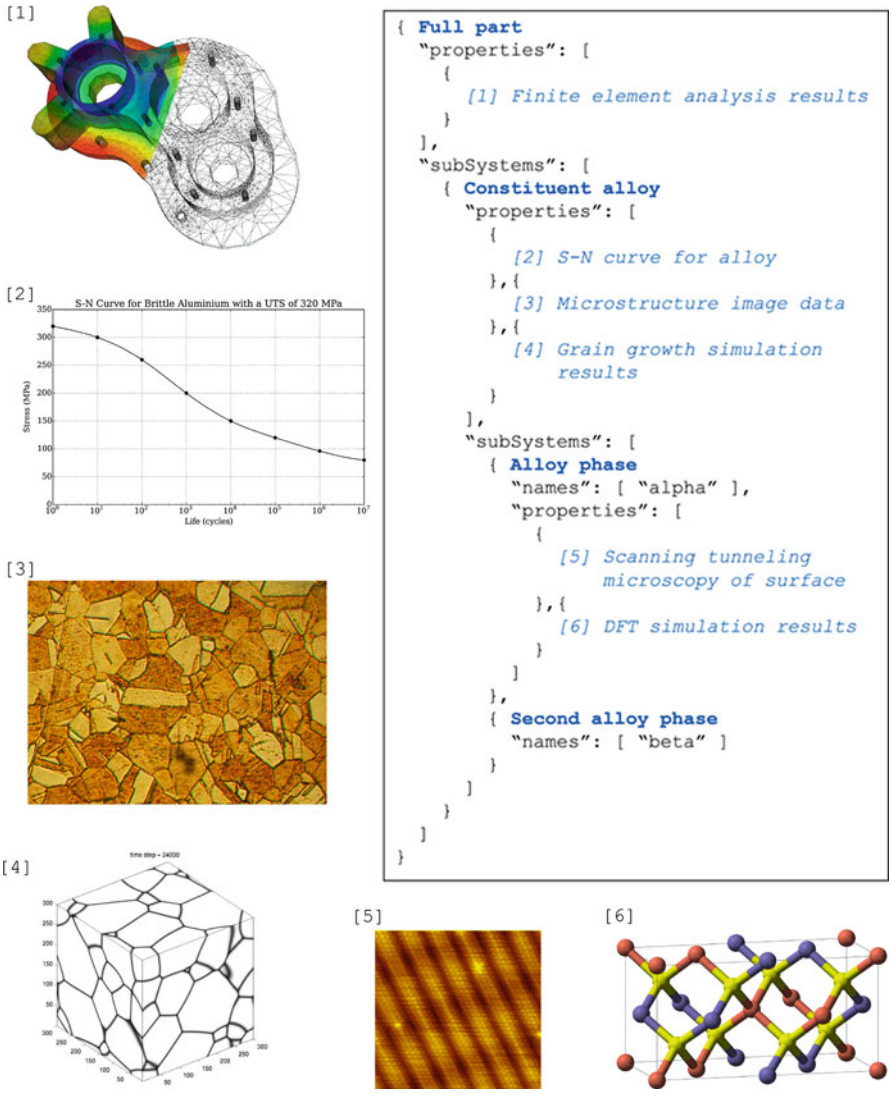

Fig. 9.1 An illustration showing how the microstructural data for a material in a physical part is stored hierarchically in the PIF from [33–38] (Copyright 2016 Materials Research Society. Reprinted with permission) Note: *S* stress, *N* number of cycles

9.1.5.1 Data Import

The Citrination platform accepts documents of all types, through upload to the web interface or through the hypertext transfer protocol (HTTP) API [39]. Once files are successfully uploaded, a copy is stored, and then the file is processed

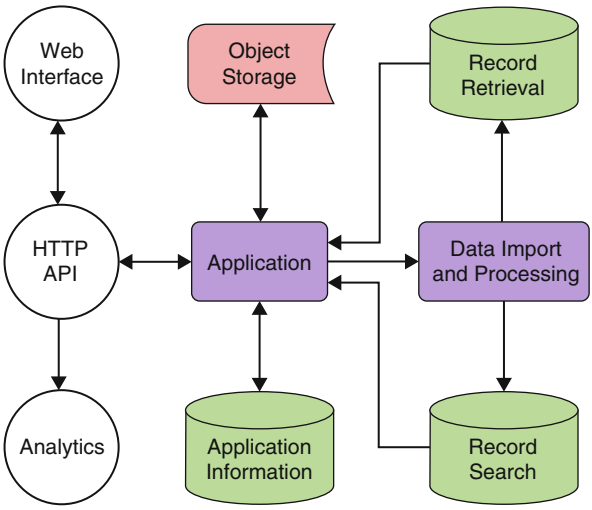

Fig. 9.2 An illustration of the data architecture used for the Citrination platform from [2] (Copyright 2016 by The Minerals, Metals & Materials Society. Used with permission). *White circles* represent entry points into the system, the *red box* represents an object storage system, *green cylinders* represent data management systems, and *purple boxes* represent application logic. The flow of data through the system is shown by the *arrows*

according to the level of support for the specific file format. The PIF is currently the only fully supported file type, and PIF files are guided through the import and processing pipelines. These records are loaded into the search and retrieval databases, allowing for users to locate and use them again in the future. Unsupported structured documents, such as spreadsheet files and experimental equipment output files, are stored so that they can be processed once support for additional file formats is incorporated on the platform. Unstructured files, such as academic journal articles, books, patents, and company documents, are the most complex to handle but often contain a wealth of valuable information. Citrine has built a variety of tools and methods for performing structured PIF extractions from unstructured documents [2]. Traditionally, extractions of this nature would be performed by human experts, but that process is time-consuming and can be error prone [40]. Automated extraction of structured information from unstructured documents is an area of active research [41].

9.1.5.2 Data Storage

As every data storage approach has pros and cons, the Citrination platform leverages several different database technologies for specialized purposes. A relational database is used to store user authentication and session information, as it provides fast response times, required for this type of data. A long response time here would make accessing the materials data unacceptably slow, as requests for this information accompany every search. This data store needs to be resilient and highly available, as it is critical to application access [2]. An object store is used to store documents as they are uploaded to the site. Object stores are commonly used in web applications, as they are considered scalable and resilient. All documents

persist in the object store and provide a ground truth in the system: at any point, these documents could be reprocessed to completely regenerate all the data in Citrination. Object stores do not offer fast response times in comparison to other technologies such as relational databases, but as the documents are accessed less frequently than the data itself, this is acceptable for the Citrination use case. The hierarchical structure of the PIF suits non-relational databases well, and the non-relational database used for the Citrination platform serves a dual purpose. It is a source of records, served up in the web interface or to the API, and it acts as a staging area for new data before they are indexed in the search engine. The search engine is required to make complex queries against large datasets, whether individual records are structured or unstructured. Lucene [42] is a popular search engine application on which Solr [43] and ElasticSearch [44] have been built. These software tools allow for scaling and deployment of Lucene on cloud-based infrastructures. A custom-built ElasticSearch plugin is used to index and query data on Citrination and allows high-level support of materials and physics language [2].

9.1.5.3 Data Access

Citrination allows access to its data through either the HTTP API [39] or a graphical web application [29]. The web application is simply an interface for the API that allows users without programming experience to navigate through the data on the system to retrieve records of interest. The web application can be used by anyone and does not require a user to create an account; however, the API uses an API key, issued only to registered users [2].

9.1.6 Data Infrastructure: Materials Data Curation System

As part of the National Institute of Standards and Technology's push to develop infrastructure to support the Materials Genome Initiative, researchers within the Material Measurement Laboratory and the Information Technology Laboratory are addressing two primary goals: (1) the materials community requires a system for exchange of materials data in community-developed machine-interoperable data formats, and (2) the materials community requires a decentralized mechanism for discovery of materials data, tools, and other resources. Two of the solutions that are being developed are the Materials Data Curation System (MDCS) and the NIST Materials Resource Registry (NMRR) software [45]. These free and open-source software projects enable the federated discovery and access of materials data and metadata.

9.1.7 Data Standards

NIST's Materials Data Curation System makes use of various XML schemas to store data in a consistent and repeatable fashion. A phase-based ontology is being developed to facilitate data curation, and existing and new XML schemas will be supported by the software. The aim of using XML schema for data storage is to create custom data formats for experimental and simulated data of different types and then to include low-level data and metadata, stored in a uniform way across focus areas in the field. This method was identified as a way to allow researchers to store their data in a format that is specific to their work while also trying to eliminate some of the challenges traditionally encountered when dealing with the typical diversity of materials data.

9.1.8 Platform Architecture

9.1.8.1 Data Import

XML templates defined to store data can also be used to generate forms that allow users to enter data and load images and other files into the MDCS. This upload is done through the web interface and allows users who may be unfamiliar with the XML format to enter data correctly when a template for their data type already exists or by using the template composer which is a graphical interface that can be used to create the required XML files. Data that is uploaded is translated into BSON and then back to XML as needed so that it is compatible with the database used for this project. Users can also upload images and other file types.

9.1.8.2 Data Storage

To manage the heterogenous data that exists in the field of materials science and engineering and to allow flexible and complex queries, the Materials Data Curation System uses a combination of NoSQL databases and relational database technologies for data storage. Data can also be harvested from other repositories that support Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), and data, metadata, and other objects are stored in XML documents in a MongoDB NoSQL database.

9.1.8.3 Data Access

For data access, APIs, web APIs, and data exchange facilities allow users to interact with the infrastructure in a number of different ways. The web interface is similar in appearance to the upload interface and is generated from the XML files.

Users can download the XML and other data files to use as necessary. Due to the transformability of XML files, there is also support to output files in other formats such as comma-separated value (CSV) files.

The NMRR, in contrast to the MDSC, does not store data but rather allows users to search across multiple different registries. This aims to allow users decentralized access to materials resources, allowing users to share data by uploading resources and metadata associated with these as well as to find data resources by using the web interface or API. The NMRR provides a list of available resources, redirecting users so that they can continue searching for data of interest [45].

9.1.9 Data Infrastructure: Materials Commons

The Materials Commons is an effort launched by a consortium of universities, national laboratories, and academic publishers. It was established in 2014 to address the goals of the Materials Genome Initiative and aims to provide open access to a broad range of materials science data, making it easier to publish and discover data. The Materials Commons has been designed to store experimental and simulation information, to be a part of the scientific workflow, to manage provenance tracking, and to provide this all in an open-source platform that allows collaboration and easy searching for data of interest.

9.1.10 Data Standards

The data model adopted by the Materials Commons focuses on storing information related to the processing-structure-property relationships. The data model stores information about samples, processes, attributes, measurements, data files, and dataset in addition to the provenance information for the data.

9.1.11 Platform Architecture

9.1.11.1 Data Import

A web interface and REST-based API have been developed to allow users to interact and add data to the repository. There is also a command line tool that allows uploads of large files or large numbers of files. Templates can be used in the web interface to record measurements with their values and provenance information. Files related to the measurements can also be uploaded.

9.1.11.2 Data Storage

Datasets are stored with metadata to facilitate discovery. They are stored securely in a system that can handle very large files. Stored files are versioned and can be grouped into datasets and projects, and files can also be shared between projects. A 390-TB Isilon cluster is used for the data storage, and this is mirrored with redundant file blocks within each mirror.

9.1.11.3 Data Access

The web interface allows search and browsing of the data in Materials Commons, and the API and command line tools allow easier downloading of large or many data files. Data is grouped into projects, and all projects are managed by a project owner who controls access and assets. Data is searchable with search provided by ElasticSearch. This search understands relationships between the different objects in the system which allows search results to be returned based on indirect matches. The system has also been designed in such a way that is easy to allow external repositories to integrate with the Materials Commons search service.

9.1.12 Data Infrastructure: Materials Data Facility

The Materials Data Facility (MDF) is a collaboration among the University of Chicago, Argonne National Laboratory, and the National Center for Supercomputing Applications and is supported by NIST and the Center for Hierarchical Materials Design (CHiMaD). The MDF aims to provide data infrastructure resources and scalable shared data services to facilitate data publication and discovery.

The Globus publication system supports the cloud-hosted services. The choice to use this service was made in an attempt to increase user adoption by providing web-based interfaces, by lowering costs, and by removing the need for management and maintenance by the end user. The functionality provided by Globus is used in conjunction with the DSpace institutional repository system.

9.1.13 Data Standards

The Materials Data Facility does not enforce structure on the data that it stores but rather allows users to add descriptive metadata to enable users to search through the data and to provide context and meaning to files they share. The metadata is arbitrary and extensible; however, collections of data can have optional or required metadata fields that are specified by administrators. The aim for this method of describing

data is that users will be able to reuse schemas and follow standards that facilitate data sharing.

9.1.14 Platform Architecture

9.1.14.1 Data Import

Users can upload data to MDF by installing the Globus endpoint, authenticating to the MDF data publication service, selecting the data to share, and then providing metadata about the source of the data, measurement conditions, etc. to facilitate with data access and reuse.

9.1.14.2 Data Storage

The MDF allows users to store datasets with their associated identification and description and allows users to specify requirements for data completeness and level of curation. Data is stored in the cloud and managed using Globus models.

9.1.14.3 Data Access

Access to the repository is possible through either the web interface or REST API interface. The uploaded data also becomes searchable using the metadata provided by the uploader. Globus data access models allow users to download and transfer files from the cloud to a user's local storage system [46].

9.2 Materials Informatics

Experiments, theory, and simulation have for decades been considered the three pillars of scientific exploration, but today data-intensive science is emerging as a fourth [47]. The field of MS&E has historically been cautious in adopting new research approaches, but the potential value of data analytics to the materials community is becoming more apparent. Materials informatics involves using algorithms to analyze large-scale materials data with the aim of providing novel insight and addressing key materials challenges [13]. The popularity of and interest in informatics is rapidly growing, and it is gaining more mainstream acceptance and visibility. Government, nonprofit, and private efforts are focusing on new ways to perform data analyses, method development, and rapid data collection, as it is believed that materials informatics will be able to accelerate the time frame for material development from invention to deployment. This is currently a process that

Table 9.2 Invention dates and commercial deployment dates of various materials

Materials technology	Year invented	Commercialization	Years (approx)	Citation
Vulcanized rubber	1839	Late 1850s	20	[50]
Low-cost aluminum	1886	Early 1900s	15	[50]
Teflon	1938	Early 1960s	25	[50]
Velcro	Early 1950s	Early 1970s	20	[50]
Polycarbonate	1953	About 1970	20	[50]
GaAs	Mid-1960s	Mid-1980s	20	[50]
GaN	1969	1993	24	[51]
NdFeB magnets	1983	Late 1980s	7	[52]
Li-ion batteries	1976	1991	15	[53]
Ferrium M54	2007	2015	8	[54]

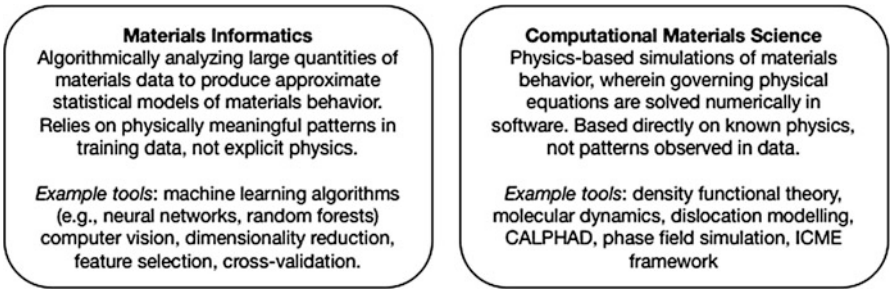

Fig. 9.3 Definitions of materials informatics and computational materials science, highlighting the distinctions we see between these broad areas of materials research (Reprinted from Ref. [3]. Copyright 2016, with permission from Elsevier)

can take from 10 to 20 years, and there has been widespread international interest in reducing this time frame [48, 49]. The time to market for a number of well-known materials is shown in Table 9.2 [32].

To date, materials informatics has not been featured as prominently in industry as physics-based computational materials methods, but it is gaining popularity, and companies such as IBM have announced their intentions to use materials informatics to assist materials discovery [55]. Materials informatics is distinctly different from computational materials science, which uses physics-based approaches such as density functional theory (DFT), molecular dynamics, or phase-field simulations to model material behavior. In contrast, materials informatics comprises a set of purely data-driven approaches that do not presuppose an understanding of underlying equations or physics principles. It is also not specific to computational researchers, as experimentalists can directly benefit from using materials informatics for modeling and data analysis [13] (Fig. 9.3).

9.2.1 Advantages of Materials Informatics

Materials informatics tools can be packaged and delivered in such a way that they do not require a user to have extensive knowledge of computational materials methods. They are also entirely empirical and thus are able to model phenomena that are not easily described by equations or neat physical descriptions. Specifically, materials informatics can be used to model phenomena that are not yet fully understood [18]. Phenomena such as corrosion and aging in alloys are two examples of properties that are extremely difficult or costly to simulate using traditional computational methods. Data-driven informatics-based models are well suited to addressing these types of questions. Informatics techniques can learn from a set of data that can be easily generalized to new materials and formulations [56–58]. In addition to this, materials informatics is typically computationally inexpensive and enables a user to abstract away important underlying physical details such as property variations as a function of crystal structure [59]. The key inputs to materials informatics are (1) sufficient training data, (2) descriptor sets that convert materials phenomena into vectors, and (3) choice of algorithm(s). Descriptor selection in particular is a crucial decision, as this choice has been shown to strongly impact the performance of models [60].

There can be benefit to using materials informatics in conjunction with traditional computational methods. Informatics can generate additional insight either by (1) using the outputs of simulation techniques as inputs for higher-level machine learning models or by (2) using the outputs of materials informatics methods as inputs to computational models when key parameters are otherwise unknown [13, 18]. Another promising area of synergy between computational materials science and materials informatics is the possibility of closed-loop, informatics-driven simulation workflows, wherein materials informatics algorithms iteratively learn from simulation outputs and then select the next simulation to run [61].

9.2.2 Applications of Materials Informatics

Materials informatics has historically focused on fundamental materials design and discovery at a laboratory scale [62], but the potential considerable practical impact exists further downstream in the materials life cycle. It is becoming increasingly clear that there must be a link between materials development, manufacturing, and life cycle. Areas such as manufacturing, research and development (R&D), and product design can all benefit from the use of materials informatics, with informatics addressing discovery, selection, and optimization for use as well as certification and manufacturing [32]. A particularly thorny problem is the gap between early-stage R&D and scale-up; a new material may have extremely promising properties in laboratory investigations but prove resistant to practical production at scale. Informatics tools can play a valuable role in coordinating knowledge across the materials life cycle to mitigate these challenges. Closely coupling theory, data, and

experiment promises to accelerate materials development and deployment [63–65]. Materials informatics in particular allows researchers to move faster and make better decisions [18]. One valuable role for informatics is assisting product developers when they consider the challenges and opportunities in the selection, manufacturing, and qualification of new materials [32].

9.2.2.1 Manufacturing

Materials informatics, which can optimize materials-related variables and help address the unique challenges organizations face in manufacturing, has an important role to play in industry. Informatics can be useful in manufacturing as it is well suited to performing key end-use analyses such as lifetime predictions (where it can correlate materials signals with product lifetime) and quality assurance (detecting the likelihood of defects occurring based on upstream data). Informatics can also be used for automatic process correction in cases where processes may drift over time and need to be adjusted or to maximize yield [18].

Smart manufacturing is an emerging area of interest that goes hand in hand with materials informatics. In smart manufacturing, devices used in the manufacturing process are connected to the Internet, and ubiquitous sensors are used to gather large amounts of data relating to environmental parameters such as temperature, pressure, flow rate, and more. These data provide an excellent foundation upon which materials informatics can be applied to gain valuable insight into the manufacturing process [60]. Integrating large-scale data collection methods into the manufacturing process in this way is becoming increasingly important as systems become more complex, and exciting but less well-understood production routes, such as additive manufacturing, are growing. The pharmaceutical industry has been aware of these processing issues for some time, as this industry must work with systems that are understood phenomenologically rather than mechanistically. These processes are often too complex to model theoretically but can be modeled using data-intensive approaches. At a later stage, if a mature mechanistic understanding is gained, then a theoretical approach can be applied to reap additional benefit [32].

9.2.2.2 Research and Development

Materials informatics can play a very important role in R&D. It is able to address many of the practical R&D requirements of companies, such as reducing the risk involved in research and development by optimizing materials selection and ensuring that new products are likely to have acceptable lifetime, are manufacturable with sufficient yield, and can be produced at scale [18]. Materials informatics can also be used to predict crystal structure [66, 67] or physical properties [68–71], to approximately model first-principle results, for materials discovery [72] and for other fundamental applications [73–75].

9.2.2.3 Product Design

Green-field discovery is important when identifying new materials with useful properties, but it is also important to look at ways to identify interesting materials from within better-known search spaces. In many cases, suitable materials for a new application have already been identified for other purposes, and it is then just a matter of identifying these candidates and repurposing them for new applications [32]. Materials informatics is useful to solve materials selection problems, allowing customers and product designers to enumerate requirements and predictively match them to suitable products and vendors. There have been a number of cases where materials targeted for one specific purpose or industry have later been used in completely unrelated products with great success. One example of this is the poly(ethylene-vinyl acetate), which is used in both the NuvaRing [76], a contraceptive, and the Croslite® foam used for structural support in Crocs™ footwear. In these cases, it is a desirable combination of properties that makes these materials suitable for the applications. Crossover successes like these are not commonplace in materials, even though exploiting commonalities in engineering requirements is becoming increasingly critical. One reason for this is that there is currently no easy way to search and analyze the properties of *all existing materials* in order to identify one that may be suitable for a new application. Materials informatics can be helpful in this regard; in addition to being able to identify candidates for new materials that could have the required performance characteristics, it can also be used to search existing materials for ones that possess the required properties.

9.2.3 Materials Informatics Limitations

As with all techniques, there are limitations to what can be achieved with materials informatics. Informatics-derived predictions will always contain error, based on the quality and quantity of underlying training data used to parameterize the informatics models. While materials informatics is well suited to identifying existing materials that may exhibit properties within a required range and predicting new chemistries that are likeliest to possess desired characteristics, it is (perhaps frustratingly) never able to prove a negative result. For example, we may be interested in asking an informatics framework to search for high-temperature superconductors, but of course we have no a priori knowledge of whether a room-temperature superconductor is at all realizable. We may search for a time and achieve no positive result, but we cannot know whether informatics has indeed satisfactorily explored the entire search space.

Further, real-world materials design work is never a simple case of maximizing a single parameter (such as T_c), as there are many property interrelationships, constraints, and trade-offs that require consideration [18]. Specifically, the Pareto front is the high-dimensional design surface over which any improvement in one material property is only achieved through a corresponding sacrifice of another

property [77, 78]. This construct implies that many desirable combinations of properties are simply unattainable. Finally, during development of new materials, it is often a challenge to determine at the outset (even with informatics) whether it is feasible to manufacture these materials at the required scale or whether the cost of the materials will be prohibitive.

9.2.4 Challenges in Materials Informatics

The principal challenge for those wishing to use materials informatics and data-intensive research approaches is data availability. Even though open-access models are becoming more popular with publishers, and government funding agencies are beginning to enforce data-sharing policies, there is still a distinct lack of clear incentives for data sharing. Further, the most common practices for storing data have led to a highly siloed, difficult-to-access materials data landscape. The majority of data available today are not structured in a way that is machine readable, and thus creating usable training sets for materials informatics can be a very time-consuming process.

Beyond data accessibility, materials informatics requires that input data are structured in a consistent, structured format that enables researchers to readily understand and repurpose others' outputs. Other fields, such as genomics, have already moved toward adopting more consistent data formats, but materials science is lagging behind in this respect. Task forces and working groups have tried to address this need on many occasions, but, as yet, there are no systems or set of standards that have been widely adopted. Citrine Informatics' JSON-based PIF format was developed in response to this issue, but its ultimate success can only be gauged by community adoption [13].

An important subtlety to the successful application of materials informatics is data diversity, specifically in terms of negative results. Informatics-based techniques "learn" the principles of materials science from exposure to many examples, not unlike human scientists. Thus, it is crucial that training data faithfully sample the true physical distributions of materials behavior. Unfortunately, the published research literature tends to focus on only the very best materials that often perform many standard deviations above the mean. Data for poorly performing materials, on the other hand, often are disregarded or forgotten. The materials science community would simultaneously benefit from greater reproducibility and unlock far greater potential in materials informatics if these negative results were widely disseminated [18].

Even in the face of these challenges, materials informatics has already been used successfully in several different areas. The case studies that follow provide more information and detail around a handful of these success stories, reinforcing the value of materials informatics in the field.

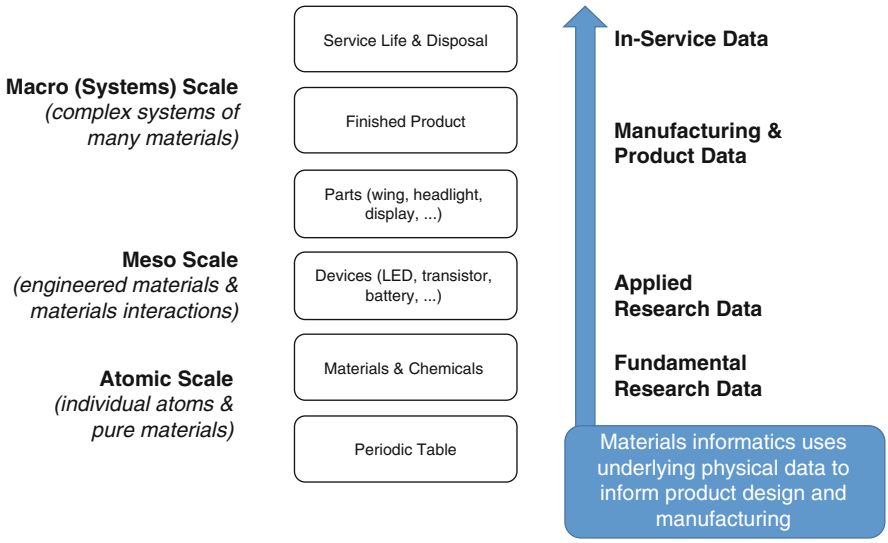

Fig. 9.4 Citrine Informatics' multiscale physical data view of product design and manufacturing. We use materials informatics techniques to analyze potentially messy data in aggregate across length scales (Reprinted from Ref. [3]. Copyright 2016, with permission from Elsevier)

9.2.5 Materials Informatics Case Studies

Citrine Informatics uses materials informatics to analyze data across length scales from the atomic to macroscale, primarily targeting industrial applications. The company specializes in analyzing materials and product behavior [18] and using underlying physical data to inform product design and manufacturing, as illustrated in Fig. 9.4 [3].

Citrine has been involved in a number of projects both in industry and academia, predominantly within materials R&D. Some example industrial use cases include vehicle lightweighting, solar materials development, formulations development, phosphor development, and more.

9.2.6 Thermoelectric Materials Discovery

In one study, machine learning drove the development of an unexpected class of thermoelectric candidates with chemical formula $RE_{12}Co_5Bi$ (RE = Gd, Er). This class of materials is quite distinct from more commonly studied structural families such as chalcogenides, skutterudites, and Zintl phases. Here, the authors used materials informatics techniques to prescreen 25,000 candidates across the key property dimensions of Seebeck coefficient, thermal conductivity, electrical

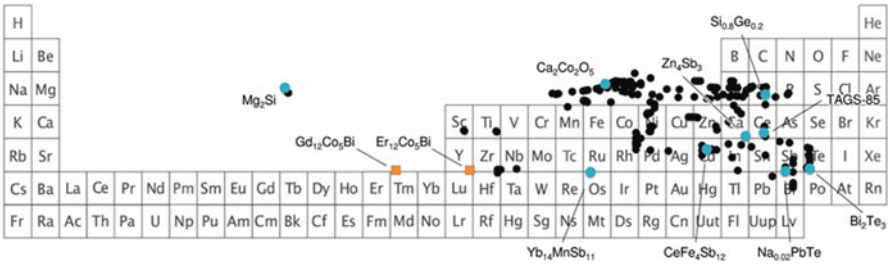

Fig. 9.5 Most known thermoelectric materials lie in a tight cluster in composition space (*black* and *blue dots*; *blue dots* have chemical formulae explicitly labeled). The recommendation engine allows the identification of new thermoelectric material families that are well outside the existing composition space of common systems in the Gaultois et al. database. In particular, we report the characterization of RE12Co5Bi (RE = Gd, Er; *orange squares*), which are chemically and structurally distinct from known thermoelectrics (Image from Ref. [59] licensed under Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/))

resistivity, and band gap and also deployed their trained models on the web (thermoelectrics.citrination.com) to allow researchers to make real-time property predictions for compounds of interest to them. Figure 9.5 demonstrates how the chemical compositions of the newly discovered materials differ substantially from current thermoelectrics. In seeking new materials with improved properties, materials researchers often restrict their search to the general neighborhood of known materials because the yield of intuition-driven green-field searches is prohibitively low; materials informatics enables scientists to circumvent this challenge and allows the identification of completely new compositions that show promising results with little overhead [59, 60].

9.2.7 Design of Polymer Dielectrics

Polymer dielectrics form essential components in applications such as electrical insulation, capacitive energy storage, organic photovoltaics, and flexible, stretchable, and wearable electronics. In terms of their dielectric or electronic potential, significant portions of the polymer chemical space remain unexplored and untapped today. This example is a demonstration of how a combination of first-principle computations and machine learning techniques led to the development of "on-demand" design models for advanced organic polymeric dielectrics.

A chemical subspace of polymers was chosen for this study, shown in Fig. 9.6. Any *n*-block polymer here is constituted of *n* of the seven basic chemical blocks, connected linearly with each other [79, 80]. A number of 4-block polymers (284 to be exact) were selected for computational data generation purposes; for each of these polymers, the three-dimensional packing and ground-state crystal structure

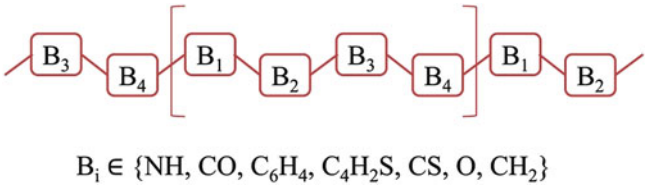

Fig. 9.6 The organic polymer chemical space selected for computations (Reproduced from Ref. [85]. Copyright © 2016 by John Wiley Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)

were predicted using the minima hopping method [81], following which density functional theory (DFT, as implemented in VASP [82]) was used to calculate the electronic band gaps (in eV, using the HSE06 functional [83]) and the dielectric constants (using density functional perturbation theory, or DFPT [84]), separated into the electronic and the ionic components. The series of steps involved in the data generation step are listed in Fig. 9.7a.

Maximizing both the dielectric constant and the band gap is necessary to improve the energy storage capacity of the polymer [79, 80, 85, 86], which is why we focus on these properties here. The total dielectric constant (ε_{total}) is given by the sum of the electronic (ε_{elec}) and the ionic (ε_{ionic}) components; the computed dielectric constants have been plotted against the computed band gaps (E_{gap}) in Fig. 9.2b. Given all these data, we next apply machine learning techniques to (a) draw correlations between crucial polymer features and the properties and (b) develop property prediction models as a function of the features [80]. Now, an essential intermediate step to performing machine learning on the computational data is to "fingerprint" the polymers, that is, to reduce them into sets of unique representative vectors: here, three kinds of chemo-structural fingerprints were used (M_I, M_{II} and M_{III}), each quantifying the types of constituent blocks and block combinations in the polymer. M_I considers the number of times each of the seven blocks appears in the polymer chain, M_{II} considers the pairs of blocks, and M_{III} considers the triplets of blocks, forming a hierarchy of polymer fingerprints that contain increasing amounts of information.

Figure 9.7c, d shows the linear correlation coefficients between the four different properties and the components of fingerprints M_I and M_{II}, respectively. This immediately enables us to identify the blocks and block pairs we are looking for in the polymer chain for a high/low dielectric constant/band gap. Dielectric polymer design rules can now be devised—for instance, more CH_2-CH_2 and CH_2-O pairs in the polymer chain will lead to higher band gaps but lower electronic dielectric constants. Whereas such a *qualitative* analysis is in itself very revealing, even more valuable is to map the polymer fingerprints to their properties using regression algorithms to develop *quantitative* predictive models. We used kernel ridge regression (KRR) [87] and the three fingerprints for this purpose to train prediction models for three properties; the best results were obtained with M_{III} and are shown in Figs. 9.8 and 9.9.

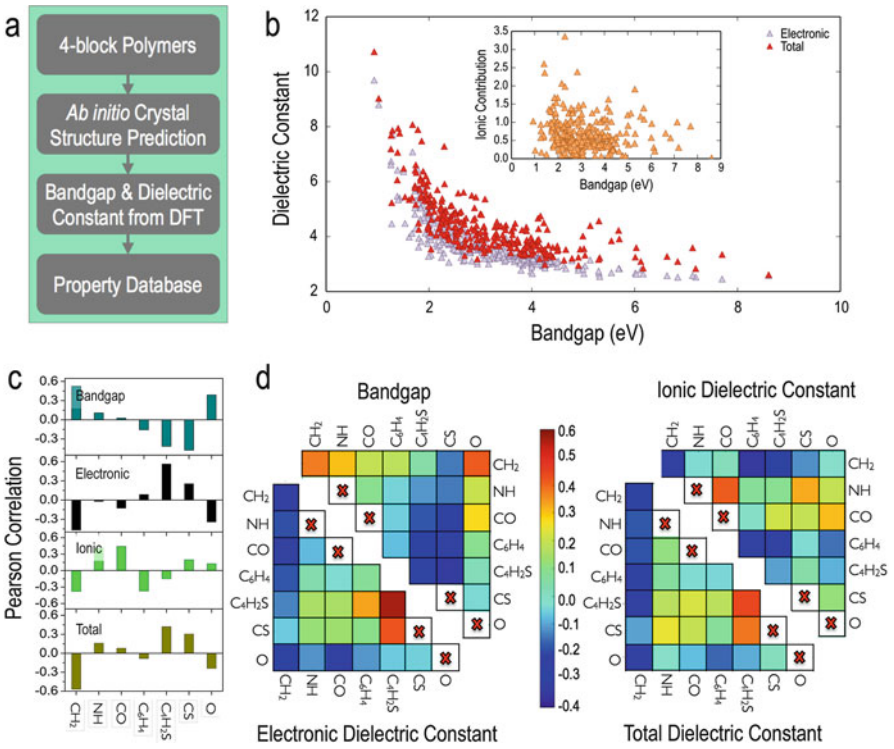

Fig. 9.7 (**a**) Steps involved in generating the computational database of four-block polymers. (**b**) DFT-computed dielectric constants (electronic, ionic, and total) plotted against the band gaps for four-block polymers. (**c**) Coefficients of linear correlation between various chemical building blocks and the properties. (**d**) Coefficients of linear correlation between various pairs of chemical building blocks and the properties (Reproduced from Ref. [80] licensed under Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/)

Around 90% of the total computational data were utilized in training the model, which was then tested on the remaining points. In the parity plots presented in Fig. 9.8a–c, the relative prediction error distributions are shown with insets, and the average errors for each property are seen to be less than 10%. This means that we have developed machine learning (ML) models which, given the fingerprint of a new polymer, instantly predict the dielectric constants and band gaps within an acceptable limit of accuracy as compared to actual DFT computations. To test the true predictive capability of the ML models, 28 random 8-block polymers were selected (given the model training was on purely 4-block polymers, this is an out-of-sample expansion) and computed their properties from DFT. Figure 9.4 shows that these values match quite well with the ML predictions. For any given population of n-block polymers, one can enumerate all the possibilities, fingerprint them, and compute the properties for all—this will enable us to populate the plot in Fig. 9.2b, leading to numerous more potential dielectric candidates.

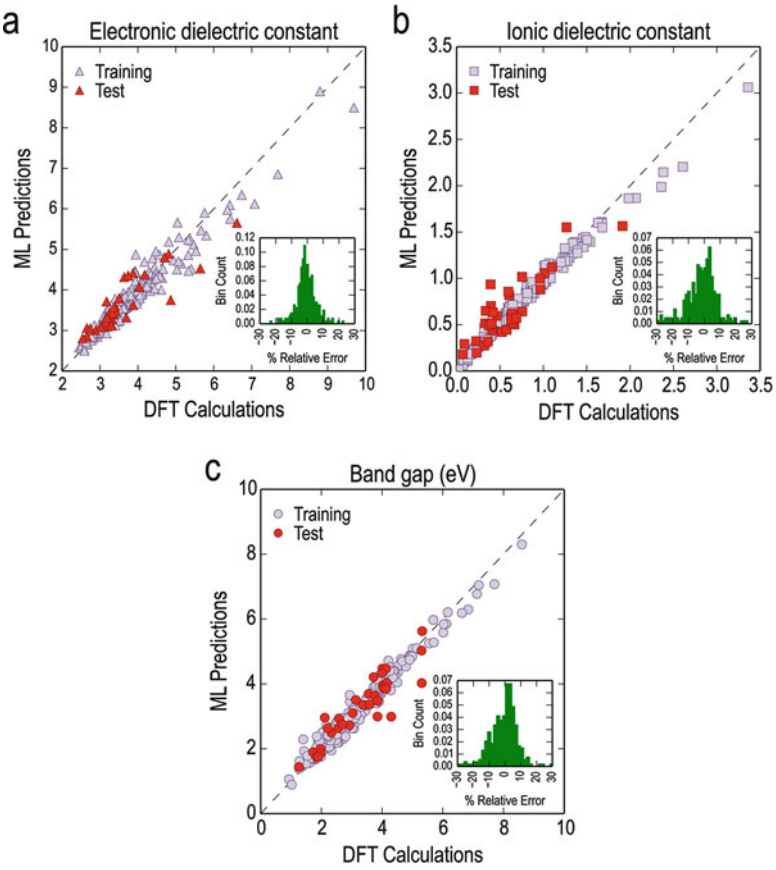

Fig. 9.8 Parity plots between ML model predictions and DFT-computed values for three properties: ε_{elec}, ε_{ionic}, and E_{gap} (Reproduced from Ref. [80] licensed under Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/))

9.2.8 Dielectric Breakdown

Predictive dielectric breakdown theories are critical to understanding the behavior and failure of dielectric insulators experiencing extreme electric fields. The intrinsic dielectric breakdown field of insulators is the theoretical limit of breakdown, determined purely by the chemistry of the material, i.e., the elements the material is composed of, the atomic-level structure, and the bonding. In this example, the intrinsic breakdown field was computed for a variety of model insulators (shown in Fig. 9.10) using laborious first principles calculations, following which machine learning schemes were used to reveal analytical relationships between the breakdown field and easily accessible material properties [88]. Such general models can guide the screening and systematic identification of high electric field-tolerant materials.

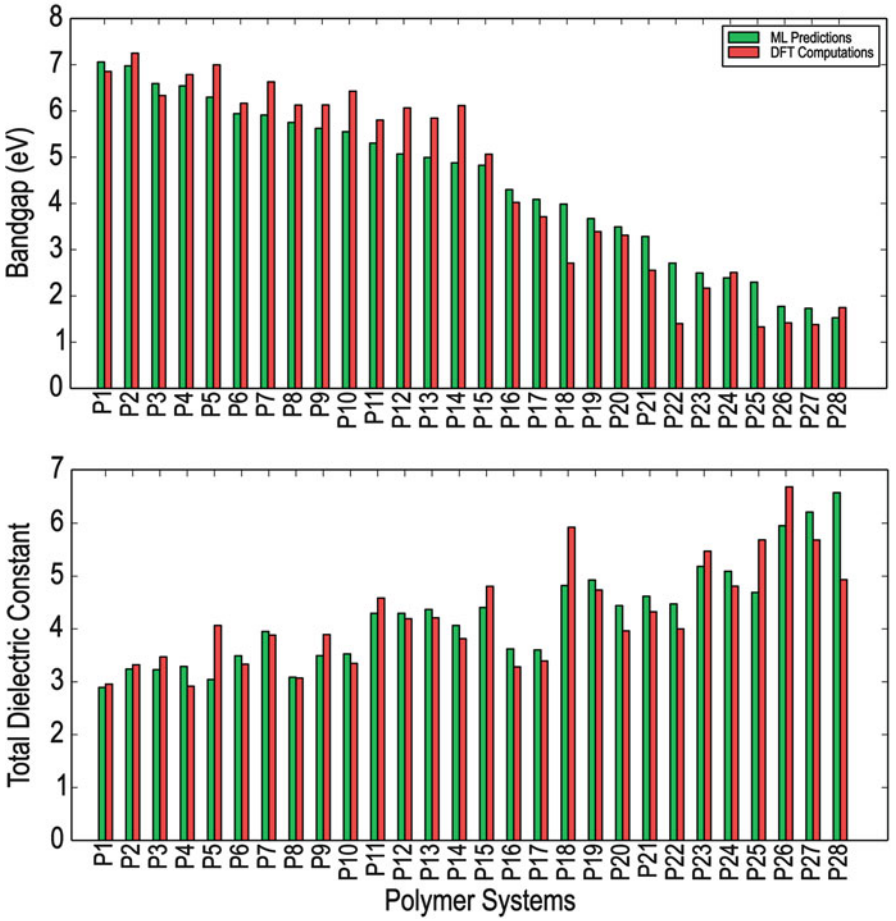

Fig. 9.9 Comparison of ML model-predicted band gaps and dielectric constants of several randomly selected eight-block polymers with their DFT-computed properties (Reproduced from Ref. [80] licensed under Creative Commons Attribution 4.0 International Public License (https://creativecommons.org/licenses/by/4.0/)

The Fröhlich-von Hippel dielectric breakdown criterion [89, 90] implemented within a first-principle density functional theory (DFT) framework [91] was used to compute the breakdown strengths of the 82 inorganic compounds. DFT was further applied to calculate the following properties to act as *primary features* for the materials, so that correlations could be drawn between them and the breakdown fields: band gap, dielectric constant, maximum and mean phonon cutoff frequency, bulk modulus, mass density, and nearest neighbor distance. The details of all the calculations (using software packages Quantum ESPRESSO and VASP) are given in ref [88]. The most important motivation for developing a phenomenological model of intrinsic dielectric breakdown is that the DFT approach to predicting this

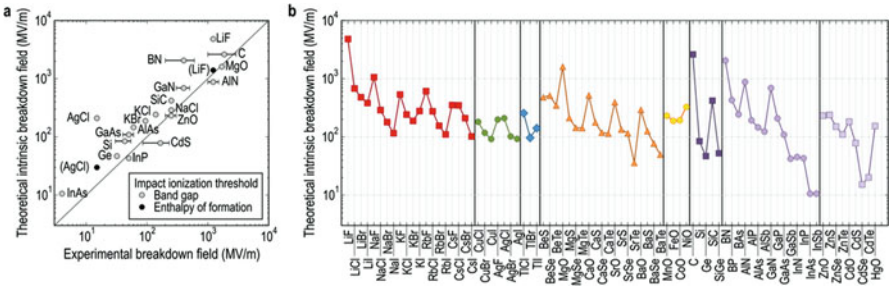

Fig. 9.10 (**a**) Comparison of the DFT-computed intrinsic dielectric breakdown field of various compounds with available experimental results (containing error bars that span the minimum and maximum known values). (**b**) DFT-computed intrinsic dielectric breakdown field for 82 reference insulators (including 79 binary compounds and 3 elemental materials (Reproduced from Ref. [88]. Reprinted with permission from Ref. [88]. Copyright 2016 American Chemical Society)

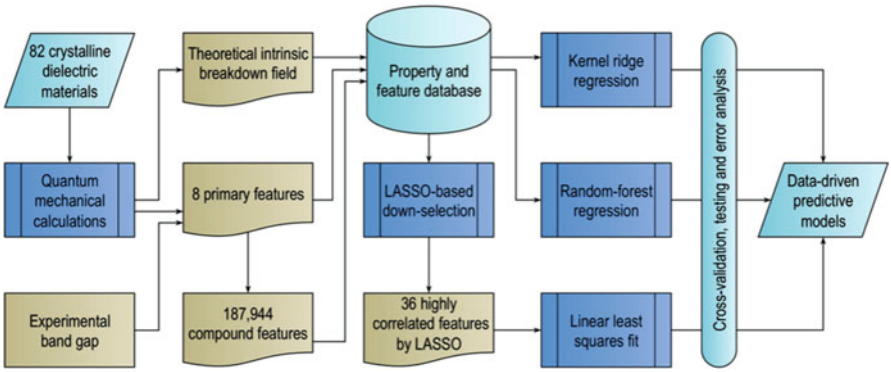

Fig. 9.11 Schematic workflow used in the data-driven discovery of a phenomenological model of intrinsic dielectric breakdown. While KRR and RFR attempt to predict the intrinsic dielectric breakdown field of a material given a set of eight primary features, the least squares pathway discovers the functional relationship between the intrinsic dielectric breakdown field and a set of compound (nonlinear) features identified by the least absolute shrinkage and selection operator (LASSO) (Reproduced from Ref. [88]. Reprinted with permission from Ref. [88]. Copyright 2016 American Chemical Society)

property, although general and accurate, is exceedingly computationally intensive even for the simple elemental or binary dielectrics (composed of just two atoms per primitive cell) considered here.

For machine learning, all the materials were converted to representative descriptors, the easiest of which were using the eight primary features. In order to capture the inherent nonlinearity in the relationships, compound features were derived out of the primary features using some prototypical functions, such as x, x^2, ln(x), or e^x. Given that this leads to nearly 200,000 compound features, a least absolute shrinkage and selection operator (LASSO)-based approach was used to extract the most important features that can act as the material descriptors to yield the best

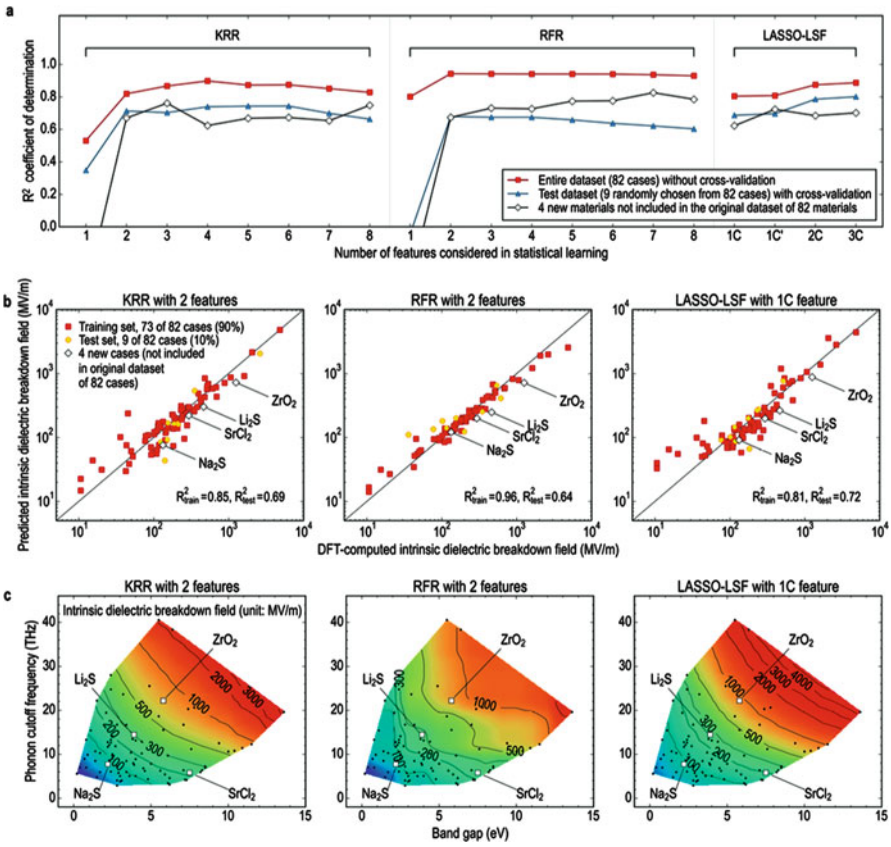

Fig. 9.12 (**a**) Coefficient of determination (R2) of different models with and without cross validation, for Ω-dimensional descriptors. While Ω ranges from one to eight in the case of KRR and RFR models, it ranges from one to three in the case of the LASSO-LSF model. R2 has also been estimated and shown for four new compounds (*Li2S*, *Na2S*, *SrCl2*, and *ZrO2*) not already included in original data set. (**b**) Parity plots comparing DFT-computed intrinsic dielectric breakdown field with the values as predicted by the KRR, RFR, and LASSO-LSF models, for the training and test sets (belonging to the original set of 82 compounds), as well as for the 4 new compounds. (**c**) Design maps for the prediction of intrinsic dielectric breakdown field using the band gap and phonon cutoff frequency. The corresponding values of these two properties of the 82 benchmark materials are indicated using *dots* and further highlighted by *shading*; the four new compounds have also been indicated (Reproduced from Ref. [88]. Reprinted with permission from Ref. [88]. Copyright 2016 American Chemical Society)

predictive models. Three kinds of learning algorithms were used here: kernel ridge regression (KRR), random forest regression (RFR), and a linear least squares fit based on the LASSO down-selected features (LASSO-LSF). The entire procedure used in this example is distilled down to a schematic workflow in Fig. 9.11, and the learning performances with the three algorithms (using different descriptor dimensions) are shown in Fig. 9.12.

The best performing 2D descriptors using both KRR and RFR involved the band gap and the phonon cutoff frequency. In the case of LASSO-based prediction, two different 1D compound descriptors were found to be equally good, and both included functions of the band gap and the phonon cutoff frequencies. Reaching a compromise between model complexity and prediction accuracy, the LASSO-LSF approach provided the following relatively simple explicit functional form for the breakdown field: $F_b = 24.442\exp(0.315\ E_g\omega_{max})$. Thus, machine learning was able to point out the two most important contributing factors to the intrinsic breakdown fields of inorganic compounds. The generality of the prediction models was tested by making predictions for four new materials not included in the learning process at all (as shown in Fig. 9.3b) and were seen to match reasonably well with the DFT-computed values. Recently, this machine learning approach was applied on a dataset of ABX_3 perovskites to compute their intrinsic dielectric breakdown strengths [92].

References

1. Westbrook, J.H., Rumble, J.R. Jr. Computerized Materials Data Systems. Gaithsburg (1983) https://www.osti.gov/scitech/biblio/6969565
2. O'Mara, J., Meredig, B., Michel, K.: Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. JOM **68**(8) 2013–2034 (2016)
3. Meredig, B.: Industrial materials informatics: analyzing large-scale data to solve applied problems in R&D, manufacturing, and supply chain. COSSMS. **21**(3), 159–166 (2016)
4. Frantzen, A., Sanders, D., Scheidtmann, J., Simon, U., Maier, W.F.: A flexible database for combinatorial and high-throughput materials science. QSAR Comb. Sci. **24**(1), 22–28 (2005)
5. Xu, Y., Yamazaki, M., Villars, P.: Inorganic materials database for exploring the nature of material. Jpn. J. Appl. Phys. **50**(11), 11RH02 (2011)
6. National Science and Technology Council Committee on Technology: Materials Genome Initiative Strategic Plan," no. June, (2014)
7. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., Persson, K.A.: Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. **1**(1), 11002 (2013)
8. Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R.H., Nelson, L.J., Hart, G.L.W., Sanvito, S., Buongiorno-Nardelli, M., Mingo, N., Levy, O.: AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. Comput. Mater. Sci. **58**, 227–235 (2012)
9. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., Wolverton, C.: Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). JOM. **65**(11), 1501–1509 (2013)
10. Holdren, J.P.: Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research. pp. 1–6, (2013)
11. Austin, T.: No Title. Mater. Discov. (2016)
12. The NoMaD Repository. [Online]. Available: http://nomad-repository.eu/cms/. Accessed: 17-Jul-2016
13. Hill, J., Mulholland, G., Pearson, K., Seshadri, R., Wolverton, C., Meredig, B.: Materials science with large scale data and informatics: unlocking new opportunities. MRS Bull. **41**, 399–409 (2016)
14. NIST Repositories.

15. Foster, I., Ananthakrishnan, R., Blaiszik, B., Chard, K., Osborn, R., Tuecke, S., Wilde, M., Wozniak, J.: Networking materials data: accelerating discovery at an experimental facility. Adv. Parallel Comput. **26**, (2015)
16. Inorganic Crystal Structure Database. [Online]. Available: https://lib.stanford.edu/inorganic-crystal-structure-database-icsd. Accessed: 09-Feb-2015
17. A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design, Acta Crystallogr. Sect. B Struct. Sci., 58, 3, 364–369,2002
18. Meredig, B.: Industrial materials informatics: analyzing large-scale data to solve applied problems in R&D, manufacturing, and supply chain. COSSMS (2016)
19. Codd, E.F.: Relational database: a practical foundation for productivity. Commun. ACM. **25**(2), 109–117 (1982)
20. Sumathi, S., Esakkirajan, S.: Fundamentals of Relational Database Management Systems
21. Sadalage, P.J., Fowler, M.: NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley, Upper Saddle River (2013)
22. Blair, J., Canon, R.S., Deslippe, J., Essiari, A., Hexemer, A., MacDowell, A.A., Parkinson, D.Y., Patton, S.J., Ramakrishnan, L., Tamura, N., Tierney, B.L., Tull, C.E.: High performance data management and analysis for tomography, p. 92121G (2014)
23. Mesnier, M., Ganger, G.R., Riedel, E.: Storage area networking - object-based storage. IEEE Commun. Mag. **41**(8), 84–90 (2003)
24. Hall, S.R., Allen, F.H., Brown, I.D.: The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Crystallogr. Sect. A Found. Crystallogr. **47**(6), 655–685 (1991)
25. Warren, J.A, Boisvert, R.F.: Building the Materials Innovation Infrastructure: Data and Standards Building the Materials Innovation Infrastructure: Data and Standards. (2012)
26. Ward, C.H., Warren, J.A., Ward, C.H.: Materials Genome Initiative : Materials Data
27. NIST Materials Data Curation System. [Online]. Available: https://mgi.nist.gov/materials-data-curation-system
28. Huck, P., Jain, A., Gunter, D., Winston, D., Persson, K.: A Community Contribution Framework for Sharing Materials Data with Materials Project. (2015)
29. Citrine Informatics, "Citrination." [Online]. Available: https://citrination.com. Accessed: 09-Feb-2015
30. Michel, K.J., Meredig, B.: Beyond bulk Single crystals: a data format for all materials structure-property-processing relationships. MRS Bull. **41**(8), 617–623 (2016)
31. Documenation of the Physical Information File (PIF) schema. [Online]. Available: http://citrineinformatics.github.io/pif-documentation/
32. Mulholland, G.J., Paradiso, S.P.: Perspective: materials informatics across the product lifecycle: selection, manufacturing, and certification. APL Mater. **4**(5), 53207 (2016)
33. No Title. [Online]. Available: https://commons.wikimedia.org/wiki/File:Elmer-pump-heatequation.png
34. No Title. [Online]. Available: https://commons.wikimedia.org/wiki/File:BrittleAluminium320MPa_S-%0AN_Curve.svg
35. No Title. [Online]. Available: https://commons.wikimedia.org/wiki/File:Microstructure_of_rolled_and_annealed_brass;_magnification_400X.jpg.
36. No Title. [Online]. Available: https://commons.wikimedia.org/wiki/File:Grgr3d_small.gif
37. No Title. [Online]. Available: https://commons.wikimedia.org/wiki/File:Atomic_resolution_Au100.JPG.
38. No Title. [Online]. Available: https://commons.wikimedia.org/wiki/File:Chalcopyrite-unit-cell-3D-balls.png
39. Citrination API Documentation
40. Seshadri, R., Sparks, T.D.: Perspective: interactive material property databases through aggregation of literature data. APL Mater. **4**(5), 53206 (2016)
41. Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., Ré, C.: Incremental knowledge base construction using DeepDive. Proc. VLDB Endow. **8**(11), 1310–1321 (2015)

42. Lucene. [Online]. Available: https://lucene.apache.org/
43. Solr. (n.a.) [Online]. Available: http://lucene.apache.org/solr
44. ElasticSearch. (n.a.) [Online]. Available: https://www.elastic.co/products/elasticsearch
45. Dima, A., Bhaskarla, S., Becker, C., Brady, M., Campbell, C., Dessauw, P., Hanisch, R., Kattner, U., Kroenlein, K., Newrock, M., Peskin, A., Plante, R., Li, S.-Y., Rigodiat, P.-F., Amaral, G. S., Trautt, Z., Schmitt, X., Warren, J., Youssef, S : Informatics infrastructure for the materials genome initiative. JOM. (2016)
46. Blaiszik, B., Chard, K., Pruyne, J., Ananthakrishnan, R., Tuecke, S., Foster, I.: The materials data facility: data services to advance materials science research. JOM. **68**(8), 2045–2052 (2016)
47. Tansley, S., Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery
48. White, A.: The materials genome initiative: one year on. MRS Bull. **37**(8), 715–716 (2012)
49. Materials in the New Millennium: National Academies Press: Washington, D.C (2001)
50. Eagar, Thomas: Bringing new materials to market. Technol. Rev. **98**(2), (1995)
51. Nakamura, S., Krames, M.R.: History of Gallium–Nitride-Based Light-Emitting Diodes for Illumination
52. Hadjipanayis, G.C., Hazelton, R.C., Lawless, K.R.: New iron-rare-earth based permanent magnet materials. Appl. Phys. Lett. **43**(8), 797 (1983)
53. Ceder, G., Whittingham, M.S., Ceder, G., Van der Ven, A., Morgan, D., Van der Ven, A., Ceder, G., Kang, B., Ceder, G., Ping Ong, S., Wang, L., Kang, B., Ceder, G., Kayyar, A., Qian, H., Luo, J., Ong, S.P., Jain, A., Hautier, G., Kang, B., Ceder, G., Reed, J., Ceder, G., Reed, J., Ceder, G.: Opportunities and challenges for first-principles materials design and applications to li battery materials. MRS Bull. **35**(9), 693–701 (2010)
54. Allison, J., Backman, D., Christodoulou, L.: Integrated computational materials engineering: a new paradigm for the global materials profession. JOM. **58**(11), 25–27 (2006)
55. Johnson, R.C.: IBM launches accelerated discovery lab. EE Times (2013)
56. Suh, C., Rajan, K., Vogel, B., Narasimhan, B., Mallapragada, S.: Informatics Methods for Combinatorial Materials Science. Wiley, Hoboken (2006)
57. Agrawal, A., Deshpande, P.D., Cecen, A., Basavarsu, G.P., Choudhary, A.N., Kalidindi, S.R.: Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. Integr. Mater. Manuf. Innov. **3**(1), 8 (2014)
58. Jee, D.-H., Kang, K.-J.: A method for optimal material selection aided with decision making theory. Mater. Des. **21**(3), 199–206 (2000)
59. Sparks, T.D., Gaultois, M.W., Oliynyk, A., Brgoch, J., Meredig, B.: Data mining our way to the next generation of thermoelectrics. Scr. Mater. (2015)
60. Gaultois, M.W., Oliynyk, A.O., Mar, A., Sparks, T.D., Mulholland, G.J., Meredig, B.: Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. APL Mater. **4**(5), 53213 (2016)
61. Peterson, A.A., Christensenb, R., Khorshidia, A.: Addressing uncertainty in atomistic machine learning. Phys. Chem. Chem. Phys. (18), 10978–10985 (2017)
62. Jain, A., Hautier, G., Moore, C.J., Ping Ong, S., Fischer, C.C., Mueller, T., Persson, K.A., Ceder, G.: A high-throughput infrastructure for density functional theory calculations. Comput. Mater. Sci. **50**(8), 2295–2310 (2011)
63. Eager, T.W.: No Title. MIT Technol. Rev. **98**(42), (1995)
64. Barnett, B., Bowen, H.K., Clark, K.: The changing paradigm for business success in advanced materials and components manufacturing. MRS Bull. **17**(4), 35–37 (1992)
65. Swink, M., Song, M.: Effects of marketing-manufacturing integration on new product development time and competitive advantage. J. Oper. Manag. **25**(1), 203–217 (2007)
66. Meredig, B., Agrawal, A., Kirklin, S., Saal, J.E., Doak, J.W., Thompson, A., Zhang, K., Choudhary, A., Wolverton, C.: Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys. Rev. B. **89**(9), 94104 (2014)
67. Faber, F., Lindmaa, A., von Lilienfeld, O.A., Armiento, R.: Crystal Structure Representations for Machine Learning Models of Formation Energies (2015)

68. Balachandran, P.V., Theiler, J., Rondinelli, J.M., Lookman, T.: Materials prediction via classification learning. Sci Rep. **5**, 13285 (2015)
69. Kong, C.S., Broderick, S.R., Jones, T.E., Loyola, C., Eberhart, M.E., Rajan, K.: Mining for elastic constants of intermetallics from the charge density landscape. Phys. B Condens. Matter. **458**, 1–7 (2015)
70. Kappes, B.B., Ciobanu, C.V.: Materials and Manufacturing Processes Materials Screening Through GPU Accelerated Topological Mapping
71. Fischer, C.C., Tibbetts, K.J., Morgan, D., Ceder, G.: Predicting crystal structure by merging data mining with quantum mechanics. Nat. Mater. **5**(8), 641–646 (2006)
72. Pyzer-Knapp, E.O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Aspuru-Guzik, A.: What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. https://doi.org/10.1146/annurev-matsci-070214-020823, (2015)
73. Isayev, O., Fourches, D., Muratov, E.N., Oses, C., Rasch, K., Tropsha, A., Curtarolo, S.: Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem. Mater. **27**(3), 735–743 (2015)
74. von Lilienfeld, O.A., Ramakrishnan, R., Rupp, M., Knoll, A.: Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties. Int. J. Quantum Chem. **115**(16), 1084–1093 (2015)
75. Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O.A., Müller, K.-R., Tkatchenko, A.: Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. J. Phys. Chem. Lett. **6**(12), 2326–2331 (2015)
76. Sarkar, N.: The combined contraceptive vaginal device (NuvaRing®): A comprehensive review. https://doi.org/10.1080/13625180500131683, (2009)
77. Sirisalee, P., Ashby, M.F., Parks, G.T., Clarkson, P.J.: Multi-criteria material selection in engineering design. Adv. Eng. Mater. **6**(12), 84–92 (2004)
78. Fonseca, C.M., Fleming, P.J.: Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization *
79. Sharma, V., Wang, C., Lorenzini, R.G., Ma, R., Zhu, Q., Sinkovits, D.W., Pilania, G., Oganov, A.R., Kumar, S., Sotzing, G.A., Boggs, S.A., Ramprasad, R.: Rational design of all organic polymer dielectrics. Nat. Commun. **5**, 4845 (2014)
80. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T.D., Lookman, T., Ramprasad, R.: Machine learning strategy for accelerated design of polymer dielectrics. Sci Rep. **6**, 20952 (2016)
81. Goedecker, S.: Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. J. Chem. Phys. **120**(21), 9911–9917 (2004)
82. Kresse, G., Hafner, J.: *Ab initio* molecular dynamics for liquid metals. Phys. Rev. B. **47**(1), 558–561 (1993)
83. Heyd, J., Scuseria, G.E., Ernzerhof, M.: Hybrid functionals based on a screened coulomb potential. J. Chem. Phys. **118**(18), 8207 (2003)
84. Baroni, S., de Gironcoli, S., Dal Corso, A., Giannozzi, P.: Phonons and related crystal properties from density-functional perturbation theory. Rev. Mod. Phys. **73**(2), 515–562 (2001)
85. Mannodi-Kanakkithodi, A., Treich, G. M., Huan, T. D., Ma, R., Tefferi, M., Cao, Y., Sotzing, G. A., Ramprasad, R.: Rational co-design of polymer dielectrics for energy storage. Adv. Mater. (2016)
86. Huan, T.D., Mannodi-Kanakkithodi, A., Kim, C., Sharma, V., Pilania, G., Ramprasad, R.: A polymer dataset for accelerated property prediction and design. Sci. Data. **3**, 160012 (2016)
87. Vu, K., Snyder, J.C., Li, L., Rupp, M., Chen, B.F., Khelif, T., Müller, K.-R., Burke, K.: Understanding kernel ridge regression: common behaviors from simple functions to density functionals. Int. J. Quantum Chem. **115**(16), 1115–1128 (2015)
88. Kim, C., Pilania, G., Ramprasad, R.: From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. Chem. Mater. **28**, 1304–1311 (2016)
89. Fröhlich, H.: Theory of dielectric breakdown. Nature. **151**(3829), 339–340 (1943)

90. Frohlich, H.: On the theory of dielectric breakdown in solids. Proc. R. Soc. A Math. Phys. Eng. Sci. **188**(1015), 521–532 (1947)
91. Sun, Y., Boggs, S.A., Ramprasad, R.: The intrinsic electrical breakdown strength of insulators from first principles. Appl. Phys. Lett. **101**(13), 132906 (2012)
92. Kim, C., Pilania, G., Ramprasad, R.: Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX_3 perovskites. J. Phys. Chem. C. **120**(27), 14575–14580 (2016)